Chapter 1

Introduction

Objective of Ch.1

To review basic methods in the organization, summarization, and description of data set

To introduce **R**, an integrated suite of software facilities for data manipulation, calculation and graphical display. [1]

To recall some useful facts of the normal and related distributions

To study one and two samples problems

[1] **R** is a free software. For details see $http://cran.stat.wisc.edu/$

1.1 Data

1.1.1 Data collection and preparation

- Four types of studies

  Experimental studies:

  1 Controlled experiments
  2 Controlled experiments with supplemental variables

  Observational studies:

  3 Confirmatory observational studies
  4 Exploratory observational studies

- Types of data

  Quantitative data

  - Continuous
  - Discrete (or Categorical)

  Qualitative (or Categorical) data

1.1.2 Descriptive statistics

Graphical methods

- Pie chart, boxplot, stem-and-leaf display, histogram, scatter plot, etc.

Airfreight breakage, Problem 1.21, page 39. A substance used in biological and medical research is shipped by airfreight to users in cartons of 1,000 ampules. The data below, involving 10 shipments, were collected on the number of times the carton was transferred from one aircraftto another over the shipment route (X) and the number of ampules found to be broken upon arrival (Y).

Table 1.1: Airfreight breakage

X_i:	1	0	2	0	3	1	0	1	2	0
Y_i:	16	9	17	12	22	13	8	15	19	11

There are a couple of methods to get this data set. For instance, you may type it as

```
> x <- c(1, 0, 2, 0, 3, 1, 0, 1, 2, 0)

> y <- c(16, 9, 17, 12, 22, 13, 8, 15, 19, 11)
```

or read it via

```
>  mydata <- read.table("CH01PR21.DAT")    #  The data is named as"mydata"
>  mydata <- read.table("A:CH01PR21.DAT")  #  For Windows
>  x <- mydata[,2]                         #  pick up 2nd column of "mydata"
>  y <- mydata[,1]                         #  pick up 1st column of "mydata"
```

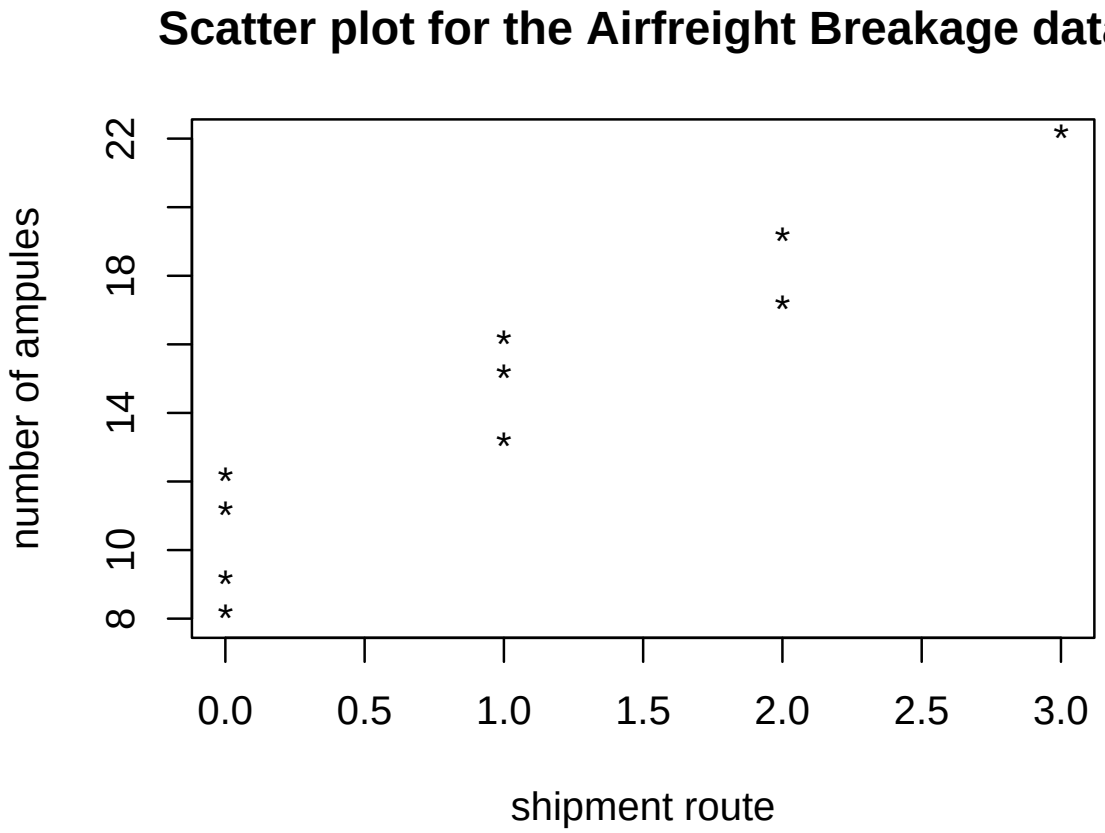
A *scatter plot* of the data set is obtained by

```
> plot(x, y)
```

Let us get a graphical impression of the distribution of Y. A `boxplot` is a graph of the five-number summary (minimum, the first or lower quartile, median, the third or upper quartile, maximum).

```
> boxplot(y)
```

Working best for small numbers of observations, a *stem-and-leaf display* gives a quick picture of the shape of a distribution while including the actual numerical values in the graph.




```
> stem(y)
   The decimal point is 1 digit(s) to the right of the |
     0 | 89
     1 | 123
     1 | 5679
     2 | 2
```

A *histogram* is a more popular graph.

```
> hist(y)                     #  histogram of Y in terms of frequency
> hist(y, freq=FALSE)         #                 in terms of relative frequency
```

Numerical methods

- Measures of central tendency

  Mean

$$\bar{x} \triangleq \frac{1}{n}\sum_{i=1}^{n} x_i$$

  Median

$$\text{median of } x \triangleq \begin{cases} x_{(\frac{n+1}{2})}, & \text{if } n \text{ is odd,} \\ \frac{x_{(\frac{n}{2})}+x_{(\frac{n}{2}+1)}}{2}, & \text{if } n \text{ is even,} \end{cases}$$

  where $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ is the ordered sample of $x_1, x_2, \ldots, x_n$.

- Measures of variation:

  Range

$$\text{range of } x \triangleq x_{(n)} - x_{(1)}$$

  Variance

$$S^2 \triangleq \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

  Standard deviation

$$S \triangleq \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Measures of relative standing:

  The 100pth quantile (or percentile)

  The 1st or lower quartile

  The 3rd or upper quartile

```
> summary(y)
  Min.   1st Qu.   Median    Mean    3rd Qu.    Max.
  8.00    11.25    14.00     14.20    16.75     22.00

> var(y)                    # sample variance of Y
 [1] 19.73333

> sd(y)                     # sample standard deviation of Y
 [1] 4.442222
```

1.2 The normal and related distributions

1.2.1 The normal distribution

Density function

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\},$$

where μ is the mean, and σ is the standard deviation (sd). In symbol, $X \sim \mathcal{N}(\mu, \sigma^2)$

Basic facts:

(i) $X \sim \mathcal{N}(\mu, \sigma^2) \implies aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$

(ii) If $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$, and X_1 and X_2 are independent, then

$$X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

(iii) If $X_1, \ldots, X_n$ is a random sample from $\mathcal{N}(\mu, \sigma^2)$, then

$$\begin{cases} \bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n}) \\ (n-1)S^2/\sigma^2 \sim \chi^2_{n-1} \\ \bar{X} \text{ and } S^2 \text{ are independent} \\ \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1} \end{cases}$$

Some useful **R** codes for the normal distribution:

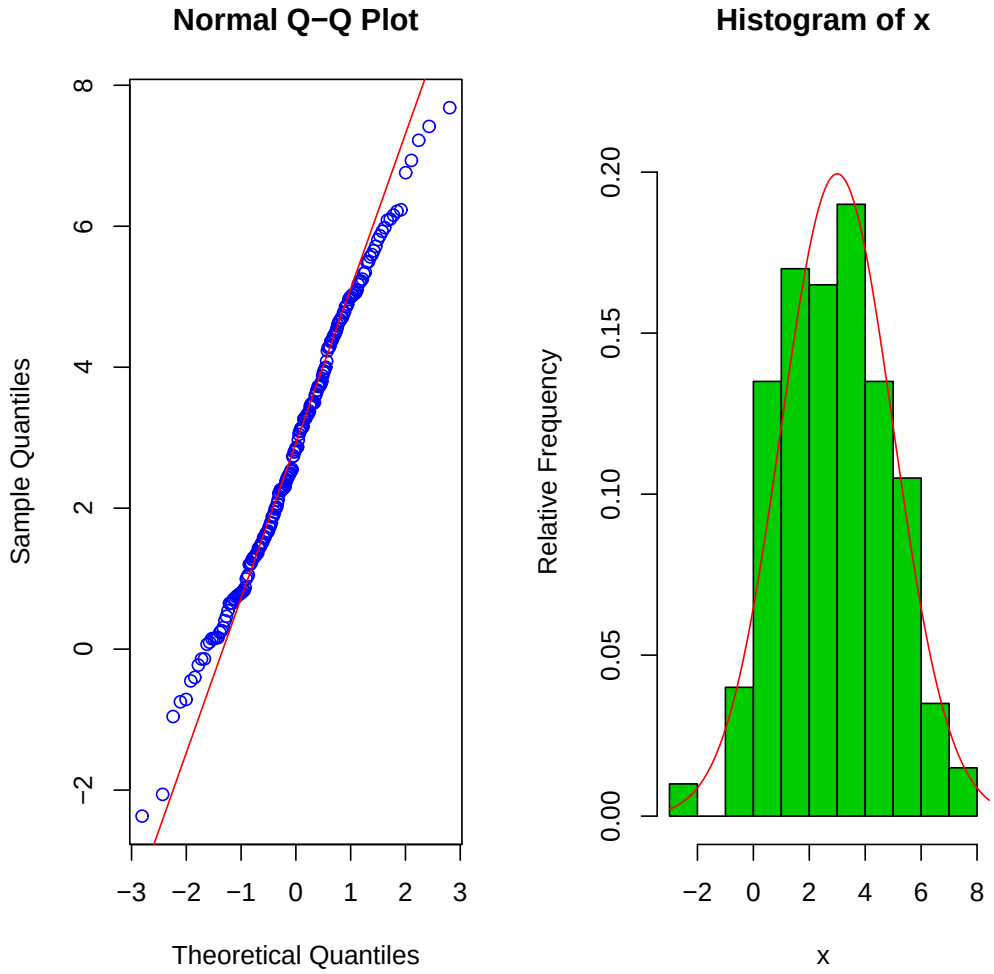
```
dnorm(x, mean, sd)       # density function

pnorm(q, mean, sd)       # cumulative distribution function (cdf)

qnorm(p, mean, sd)       # quantile function

rnorm(m, mean, sd)       # generate m normal random numbers
```

The *Normal Q-Q plot* or *Normal probability plot* is often used to check whether a data set comes from a normal distribution.

```
> qqnorm(x)                # Normal Q-Q plot

> qqline(x)                # add a straight line to the normal Q-Q plot
```

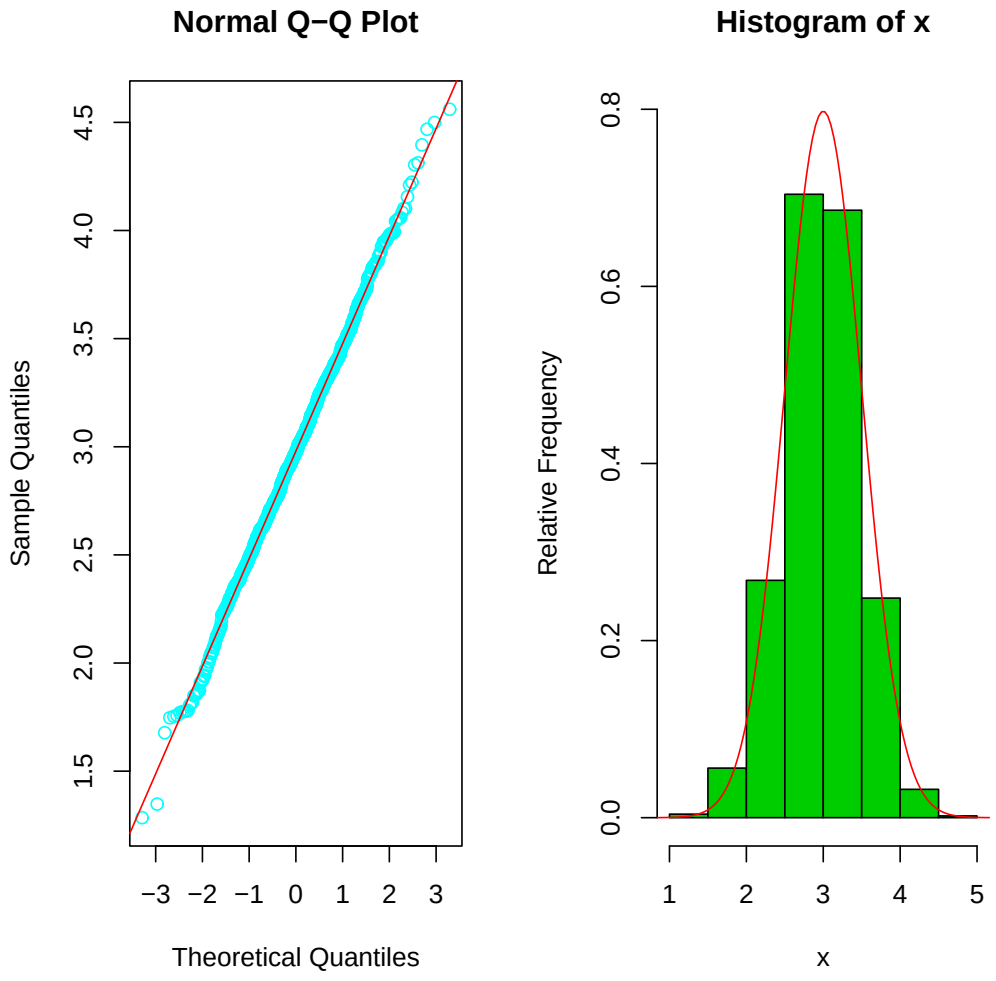
Example 1. Generate 200 normal random numbers, and then check the normality via various plots.

```
> x <- rnorm(200,3,2)    # generate 200 normal random numbers with mean=3, sd=2

> qqnorm(x)              # check the normality
> qqline(x, col=2)

> hist(x, freq=FALSE)    # histogram,  or
                         # > hist(x, freq=FALSE, ylim=c(0,0.28))

> lines(seq(-3, 9, length=200), dnorm(seq(-3, 9, length=200), 3,2), col=2)
                         # add the normal density curve (mean=3,sd=2)
```



Example 2. Generate two groups of normal random numbers, and then look at their sum.

```
> x1 <- rnorm(1000,1,0.4)   # generate 1000 normal random numbers with mean=1, sd=0.4
> x2 <- rnorm(1000,2,0.3)   # generate 1000 normal random numbers with mean=2, sd=0.3
> x <- x1+x2                # the sum of x1 and x2

> qqnorm(x)
> qqline(x, col=3)

> hist(x, freq=FALSE, ylim=c(0,0.8))
> lines(seq(-1, 6, length=200), dnorm(seq(-1, 6, length=200), 3,0.5), col=2)
                            # add the normal density curve (mean=3, sd=0.5)
```



- Correlation test & Q-Q plot for normality

Consider a set of observations $x_1, \ldots, x_n$. Denote the ordered sample as $x_{1:n} \leq \ldots \leq x_{n:n}$.

An approximation of the expected value of the kth smallest order statistic under normality is

$$sd(x) \cdot \Phi^{-1}(\frac{k - 0.375}{n + 0.25}),$$

where $\Phi^{-1}(t)$ is the quantile function (or inverse function) of the standard normal distribution.

The correlation test is based on the coefficient of correlation between the ordered sample and their expected values under normality.

```
> x <- rnorm(50, 1, 2)
> x.order <- sort(x)            # ordered sample of x
```

Their expected values are

```
> x.exp <- sd(x)*qnorm((1:length(x)-0.375)/(length(x)+0.25))
```

The coefficient of correlation between the ordered sample and their expected values under normality is

```
> cor(x.order, x.exp)
  [1] 0.9909029
```

For the level of significance $\alpha = 5\%$, the critical value for $n = 50$ is 0.977 (Table B.6). Since $0.9909 > 0.977$, we conclude that the distribution of x doesn't depart substantially from a normal distribution.

Compare the normal Q-Q plot of x with the plot of `x.order` against `x.exp`. What do you see?

```
> par(mfrow=c(2,1))
> plot(x.exp, x.order)
> qqnorm(x)
```

1.2.2 Chi-square distribution

A random variable X is said to have a *chi-square* distribution with k degrees of freedom, if it has the density

$$f(x) = \frac{1}{\Gamma(k/2)2^{k/2}} x^{k/2-1} \exp\{-\frac{1}{2}x\}, \qquad x > 0,$$

where k is called the number of degrees of freedom (df). In symbol, $X \sim \chi^2_k$

◇ Some useful codes:

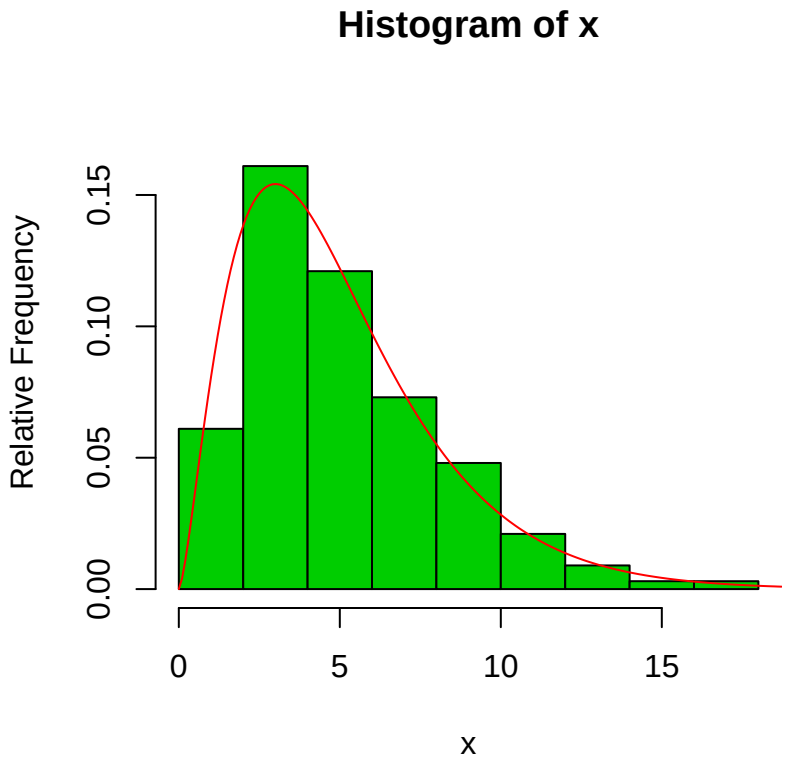
```
dchisq(x, df)            # density function

pchisq(q, df)            # cumulative distribution function (cdf)

qchisq(p, df)            # quantile function

rchisq(m, df)            # generate m chi-square random numbers
```

◇ Graphs for the density and cdf of Chi-square (df=k)

```
k <-                             # select the df

x <- seq(0, 10, length=5000)     # 0<x<10

density <- dchisq(x,k)           # density

plot(x, density, type="b", xlab="", ylab="density", col=2)

cdf <- pchisq(x, k)              # cdf

plot(x, cdf, type="s", xlab="", ylab="cdf", col=3)
```

Theorem. If $X_1, \ldots, X_k \overset{\text{i.i.d}}{\sim} \mathcal{N}(0,1)$, then $\sum_{i=1}^{k} X_i^2 \sim \chi_k^2$.

This theorem presents the relationship between the normal and chi-square distributions. To see this, we generate 5 groups of normal random numbers, and then compare their squared sum with χ_5^2.




```
> x1 <- rnorm(500,0,1)   # generate 500 standard normal random numbers
> x2 <- rnorm(500)       # same as  rnorm(500,0,1)
> x3 <- rnorm(500)
> x4 <- rnorm(500)
> x5 <- rnorm(500)
> x <- x1^2+x2^2+x3^2+x4^2+x5^2
> hist(x, freq=FALSE)    # or > hist(x, freq=FALSE, ylim=c(0,0.18), col=3)

> lines(seq(0, 20, length=200), dchisq(seq(0, 20, length=200), 5), col=2)
                         # add the chi-square density curve (df=5)
```

1.2.3 Student's t distribution

A random variable X is said to have a *Student's t* distribution with k degrees of freedom, if it has the density

$$f(x) = \frac{\Gamma((k+1)/2)}{\sqrt{k\pi}\Gamma(k/2)}(1+\frac{x^2}{k})^{-(k+1)/2}, \qquad -\infty < x < \infty,$$

where k is called the number of degrees of freedom (df). In symbol, $X \sim t_k$

◇ Some useful codes:

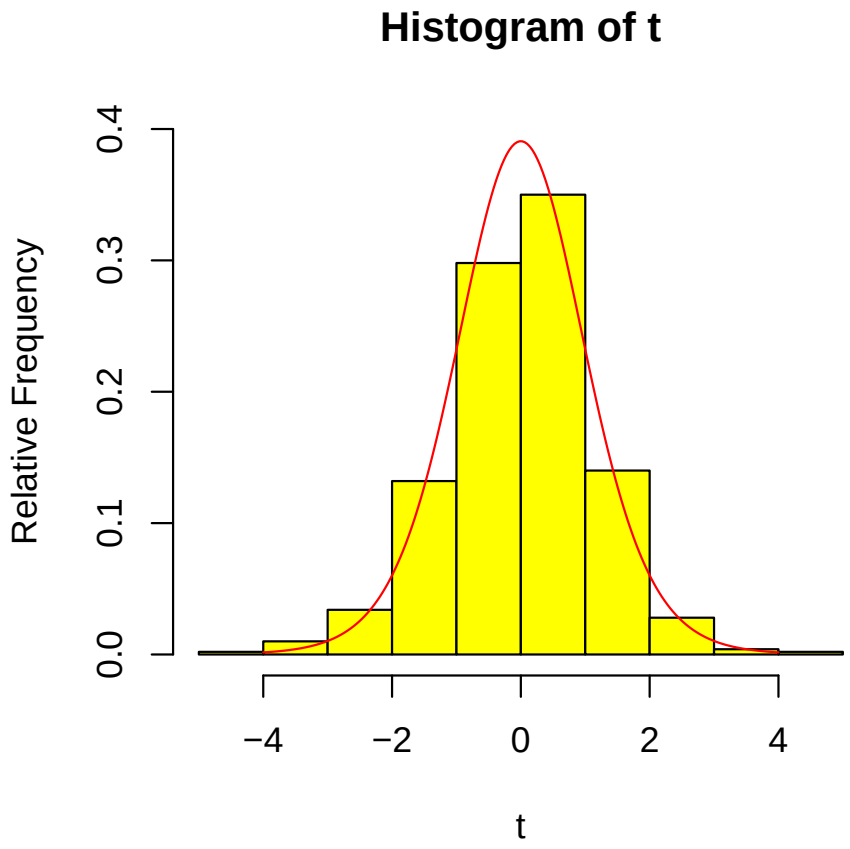
```
dt(x, df)           # density function

pt(q, df)           # cumulative distribution function (cdf)

qt(p, df)           # quantile function

rt(m, df)           # generate m t  random numbers
```

◇ Graphs for the density and cdf of Student's t (df=k)

```
k <-                              # select the df

x <- seq(-4, 4, length=5000)

density <- dt(x,k)                # density

plot(x, density, type="b", xlab="", ylab="density", col=2)

cdf <- pt(x, k)                # cdf

plot(x, cdf, type="s", xlab="", ylab="cdf", col=3)
```

Theorem. If X and Y are independent, $X \sim \mathcal{N}(0,1)$ and $Y \sim \chi^2_k$, then $\frac{X}{\sqrt{Y/k}} \sim t_k$.

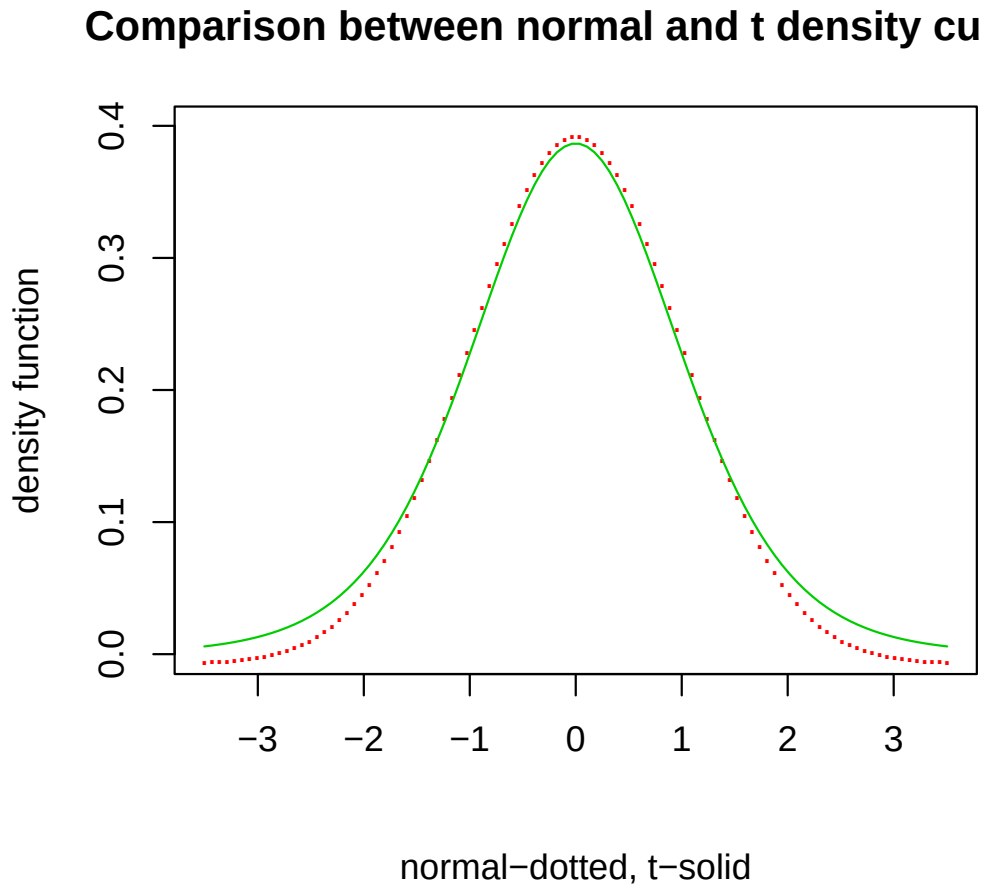
This theorem presents the relationship among Student's t, normal and chi-square distributions.




```
> x <- rnorm(500,0,1)   # generate 500 standard normal random numbers
> y <- rchisq(500,12)
> t <- x/sqrt(y/12)

> hist(t, freq=FALSE, ylim=c(0,0.4))

> lines(seq(-4, 4, length=200), dt(seq(-4, 4, length=200), 12), col=2)
                        # add the t density curve (df=12)
```

Comparison between normal and t density curves

```
> bounds <- range(qnorm(c(0.05,0.95), qt(c(0.05,0.95),8)))
                                        #  cover 5% - 95%

> x <- seq(bounds[1],bounds[2],length=100)


> plot(x, dnorm(x, 0, 1),xlim=bounds,ylim=range(c(dnorm(x, 0,1), dt(x,8))),
      col=2, pch=".", xlab="", ylab="density function", sub="normal-dotted, t-solid",
      main="Comparison between normal and t density curves")
                                         #  Standard normal density curve

> lines(x, dt(x, 8), col=3, pch="*")      #  Student's  t with df=8
```



1.2.4 F distribution

A random variable X is said to have an F distribution with degrees of freedom k_1 and k_2, if it has the density

$$f(x) = \frac{\Gamma((k_1+k_2)/2)(k1/k2)^{k1/2}}{\Gamma(k_1/2)\Gamma(k_2/2)} \frac{x^{k_1/2-1}}{(1+k_1x/k_2)^{(k_1+k_2)/2}}, \qquad x > 0.$$

In symbol, $X \sim F_{k_1,k_2}$.

The order of k_1 and k_2 is important, since $1/X \sim F_{k_2,k_1}$ whenever $X \sim F_{k_1,k_2}$.

◇ Some useful codes:

```
df(x, df1, df2)            # density function

pf(q, df1, df2)            # cumulative distribution function (cdf)

qf(p, df1, df2)            # quantile function
```

◇ Graphs for the density and cdf of F

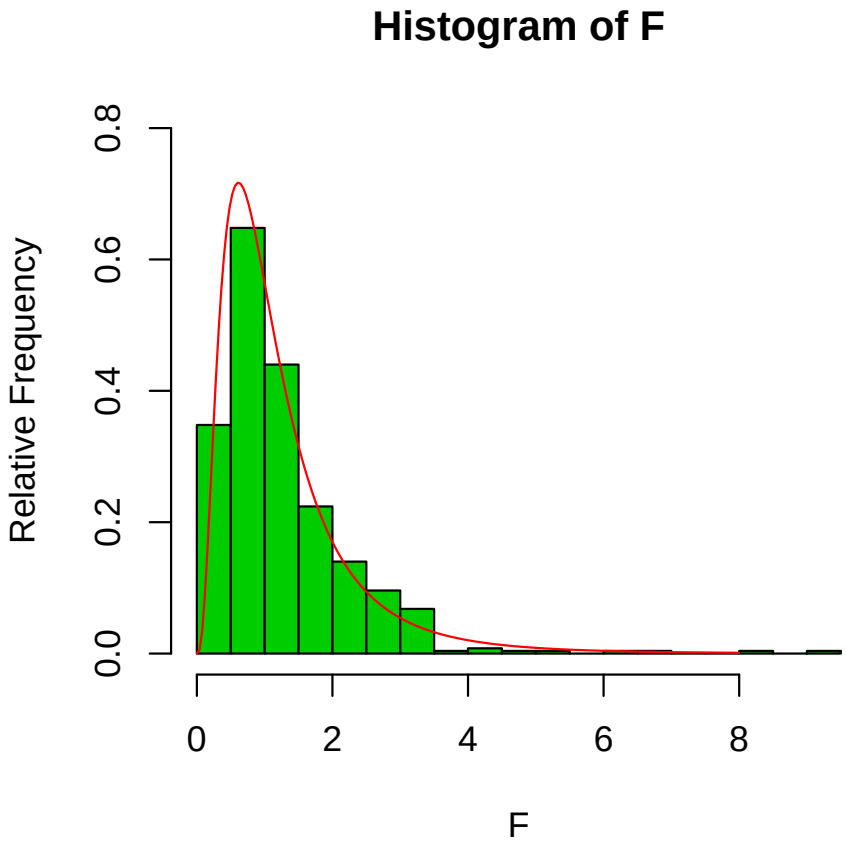
```
k1 <-                              # select the df
k2 <-

x <- seq(0, 8, length=5000)

density <- df(x, df1, df2)         # density

plot(x, density, type="b", xlab="", ylab="density", col=2)

cdf <- pf(q, df1, df2)             # cdf

plot(x, cdf, type="s", xlab="", ylab="cdf", col=3)
```

The following presents the relationships among F, Student's t and chi-square distributions.

Theorem.

(1) If X and Y are independent, $X \sim \chi^2_{k_1}$ and $Y \sim \chi^2_{k_2}$, then $\frac{X/k_1}{Y/k_2} \sim F_{k_1,k_2}$.

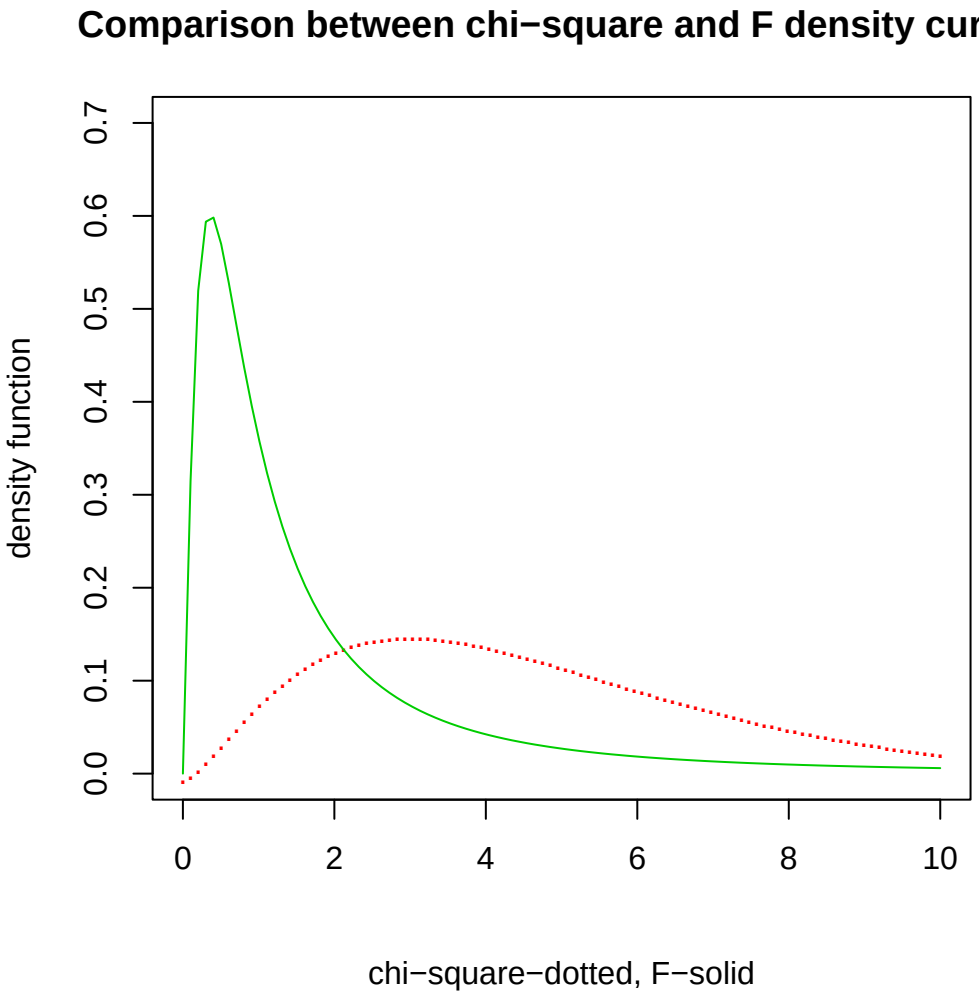
(2) If $T \sim t_k$, then $T^2 \sim F_{1,k}$.




```
> x <- rchisq(500,8)   # generate 500 chi-square random numbers
> y <- rchisq(500,9)
> F <- (x/8)/(y/9)
> hist(F, freq=FALSE)  # or hist(F, freq=FALSE, ylim=c(0, 0.8), br=20)

> lines(seq(0, 8, length=200), df(seq(0, 8, length=200), 8,9),col=2)
                       # add the F density curve (df1=8, df2=9)
```

Comparison between chi-square and F density curves

```
> x <- seq(0, 10,length=100)

> plot(x, dchisq(x, 5), ylim=c(0, 0.7),
      col=2, pch=".", xlab="", ylab="density function", sub="chi-square-dotted, F-solid",
      main="Comparison between chi-square and F density curves")
                                        #  chi-square density curve

> lines(x, df(x, 5,3), col=3, pch="F")      # F density curve
```



1.3 One sample

1.3.1 The likelihood principle

The basic idea is that *only* the actual observed data x should be relevant to making conclusions or evidence about parameter θ.

The key concept in the likelihood principle is that of the likelihood function.

As an example, consider a normal population with unknown mean μ and variance σ^2. Based on a set of observations $x_1, \ldots, x_n$, the likelihood function of parameters is

$$\begin{aligned} L(\mu, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{1}{2\sigma^2}(x_i - \mu)^2\} \\ &= \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\} \end{aligned}$$

To estimate μ and σ^2, a common and reasonable approach is the maximum likelihood method. The maximum likelihood estimators $\hat{\mu}$ and $\hat{\sigma}^2$ of μ and σ^2 are found by maximizing $L(\mu, \sigma^2)$. Clearly,

$$\hat{\mu} = \bar{x}, \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{n-1}{n}s^2$$

To derive tests concerning μ and/or σ^2, a wildly used method is the likelihood ratio test, which is based on likelihood functions.

1.3.2 One sample (from a single population), known variance

Assumptions:

(i) A random sample $X_1, \ldots, X_n$ from $\mathcal{N}(\mu, \sigma_0^2)$;

(ii) σ_0^2 is known.

Hypothesis testing problem:

$$\mathcal{H}_0 : \mu = \mu_0 \quad (\mu_0 \text{ is a given constant}) \qquad \longleftrightarrow \qquad \mathcal{H}_a : \mu \neq \mu_0 \text{ (two-sided)}$$

Test statistic:

$$U \triangleq \frac{\bar{X} - \mu_0}{\sigma_0/\sqrt{n}}$$

When $\mathcal{H}_0$ is true, $U \sim \mathcal{N}(0, 1)$.

Two ways to make a decision

(i) When controlling the level of significance at α, the decision rule is

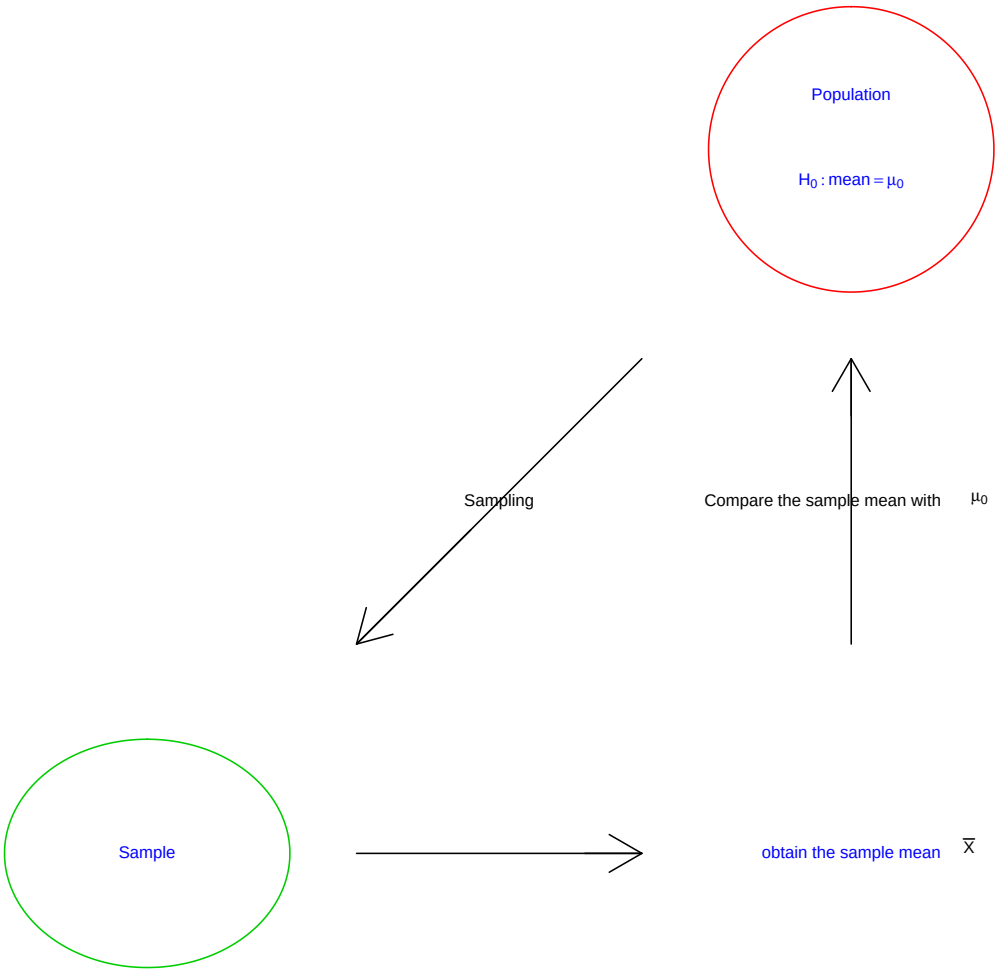
$$\begin{cases} \text{If } |observed\ U| \leq qnorm(1-\alpha/2), & \text{conclude } \mathcal{H}_0 \\ \text{If } |observed\ U| > qnorm(1-\alpha/2), & \text{conclude } \mathcal{H}_a \end{cases}$$

(ii) Report $p - value$, which may be obtained via

$$p - value \triangleq 2 \times \{1 - pnorm(\ abs(observed\ U)\ )\}$$

A smaller $p - value$ leads us to conclude $\mathcal{H}_a$.

Intuitive idea



The null hypothesis $\mathcal{H}_0 : \mu = \mu_0$ says that the population mean is μ_0. If this is the case, the sample mean $\bar{x}$ of observations $x_1, \ldots, x_n$ has to be close to μ_0 enough, because `the sample is drawn from the population`. Equivalently, $|\bar{x} - \mu_0|$ or U has to be close to 0 enough. The question is then: how close is "close enough "? For this, a threshold is needed and the decision rule is required.

When $\mathcal{H}_0$ is true, $U \sim \mathcal{N}(0, 1)$. The decision rule with this test statistic when controlling the level of significance at α is

$$\begin{cases} \text{If } |observed\ U| \leq qnorm(1 - \alpha/2), & \text{conclude } \mathcal{H}_0 \\ \text{If } |observed\ U| > qnorm(1 - \alpha/2), & \text{conclude } \mathcal{H}_a \end{cases}$$

- **Derive the U-test via the likelihood-ratio-test method**

The whole parameter space in this case is $\Theta = \{\mu \in (-\infty, \infty)\}$;

The parameter space associated with $\mathcal{H}_0$ is $\Theta_0 = \{\mu_0\}$.

The likelihood function of parameter μ is

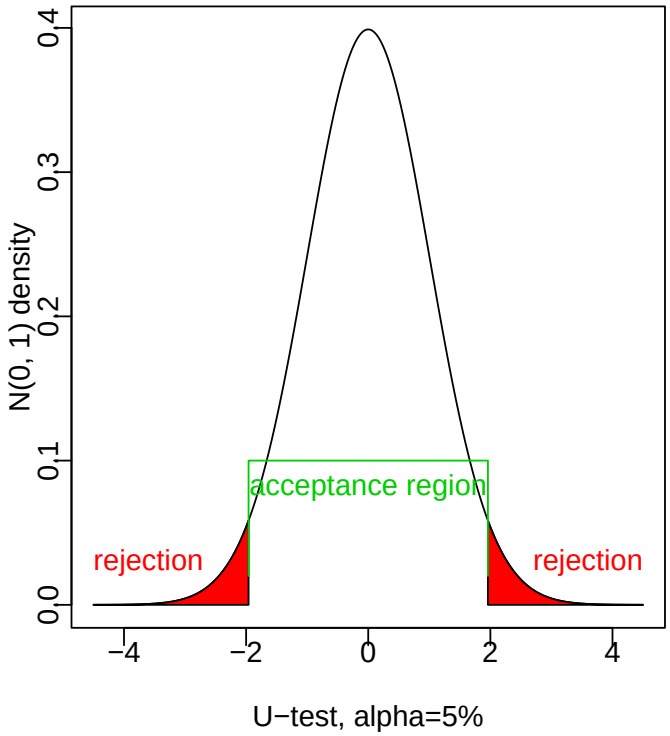
$$\begin{aligned} L(\mu) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_0} \exp\{-\frac{1}{2\sigma_0^2}(x_i - \mu)^2\} \\ &= \frac{1}{(\sqrt{2\pi}\sigma_0)^n} \exp\{-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(x_i - \mu)^2\}. \end{aligned}$$

The likelihood ratio is given by

$$\begin{aligned} \Lambda &\triangleq \frac{\sup_{\Theta_0} L(\mu)}{\sup_{\Theta} L(\mu)} \\ &= \frac{L(\mu_0)}{\sup_{\Theta} L(\mu)} \\ &= \frac{L(\mu_0)}{L(\bar{x})} \qquad \text{(why?)} \\ &= \frac{\exp\{-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(x_i - \mu_0)^2\}}{\exp\{-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(x_i - \bar{x})^2\}} \\ &= \exp\{-\frac{n}{2\sigma_0^2}(\bar{x} - \mu_0)^2\} \\ &= \exp\{-\frac{U^2}{2}\}, \end{aligned}$$

which is a decreasing function of $|U|$. Note that a smaller Λ would lead us to conclude $\mathcal{H}_a$. (why?) Equivalently, a larger $|U|$ would lead us to conclude $\mathcal{H}_a$.

Example Generate 50 normal random numbers with mean=3 and sd=2. Assume $\sigma_0 = 2$ and Test $\mathcal{H}_0 : \mu = 3.5 \quad \longleftrightarrow \quad \mathcal{H}_a : \mu \neq 3.5$




```
>  x <- rnorm(50, 3, 2)
>  u <- (mean(x)-3.5)/(2/sqrt(length(x)))
    [1] -3.169392                 #  observed U value
>  2*(1-pnorm(abs(u)))           #  p-value
    [1] 0.001527584               #  What's your conclusion?
```

Let $\alpha = 5\%$. To evaluate the acceptance region, use

```
> acceptance <- c(-qnorm(1-0.05/2), qnorm(1-0.05/2))
> acceptance
  [1] -1.959964  1.959964
```

The observed U-value is -3.169392, in the rejection region. So conclude H_a.

- **Pivotal - quantity method** for finding a confidence interval

$$\begin{cases} \text{A } \textit{statistic} & \triangleq & \text{a function of sample } X_1,\ldots,X_n \\ & & \text{that does NOT depend on any parameter } \theta \\ & & \text{(Its distribution may depend on } \theta) \\ \text{A } \textit{pivotal quantity} & \triangleq & \text{its distribution does NOT depend on parameter} \\ & & \text{(It may be a function of both } X_1,\ldots,X_n \text{ and } \theta) \end{cases}$$

As an illustration, consider a random sample $X_1,\ldots,X_n$ from $\mathcal{N}(\mu,\sigma_0^2)$, where μ is a parameter and σ_0^2 a constant. In this case,

$$\begin{cases} \bar{X}\sim\mathcal{N}(\mu,\frac{\sigma_0^2}{n}) & \text{(a statistic)} \\ \frac{\bar{X}-\mu}{\sigma_0/\sqrt{n}}\sim\mathcal{N}(0,1) & \text{(a pivotal quantity)} \end{cases}$$

For a given $0<\alpha<1$, using the standard normal table or `qnorm(1-alpha/2)` find q such that

$$P\Big(\Big|\frac{\bar{X}-\mu}{\sigma_0/\sqrt{n}}\Big|\le q\Big)=1-\alpha.$$

From this we form a $100(1-\alpha)\%$ confidence interval for μ,

$$(\bar{X}-q\cdot\frac{\sigma_0}{\sqrt{n}},\ \bar{X}+q\cdot\frac{\sigma_0}{\sqrt{n}})$$

Question 1. What are you going to do if σ_0^2 above is a unknown parameter? which is called a nuisance parameter.

Question 2. Consider a random sample $X_1,\ldots,X_n$ from $\mathcal{N}(\mu,\sigma^2)$, and use the pivotal-quantity method to form a $100(1-\alpha)\%$ confidence interval for σ^2 under the condition

(i) μ is a constant;

(ii) μ is an unknown parameter.

Question 3. Is there a relationship between an acceptance region and a confidence interval?

1.3.3 One sample (from a single population), unknown variance

Assumptions:

(i) A random sample $X_1, \ldots, X_n$ from $\mathcal{N}(\mu, \sigma^2)$;

(ii) σ^2 is unknown.

Hypothesis testing problem:

$$\mathcal{H}_0 : \mu = \mu_0 \quad (\mu_0 \text{ is a given constant}) \qquad \longleftrightarrow \qquad \mathcal{H}_a : \mu \neq \mu_0 \text{ (two-sided)}$$

Test statistic:

$$T \triangleq \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \qquad \text{(Student's t test)}$$

When $\mathcal{H}_0$ is true, $T \sim t_{n-1}$.

Two ways to make a decision

(i) When controlling the level of significance at α, the decision rule is

$$\begin{cases} \text{If } |observed\ T| \leq qt(1-\alpha/2, n-1), & \text{conclude } \mathcal{H}_0 \\ \text{If } |observed\ T| > qt(1-\alpha/2, n-1), & \text{conclude } \mathcal{H}_a \end{cases}$$

(ii) Report $p - value$, which may be obtained via

$$p - value \triangleq 2 \times \{1 - pt(\ abs(observed\ T),\ n-1)\}$$

A smaller $p - value$ leads us to conclude $\mathcal{H}_a$.

Question: How to make a decision for one-sided test problem

$$\mathcal{H}_0 : \mu = \mu_0 \quad (\mu_0 \text{ is a given constant}) \qquad \longleftrightarrow \qquad \mathcal{H}_a : \mu > \mu_0 \text{ (one-sided)}$$

Derive Student's t test via the likelihood-ratio-test method

The whole parameter space in this case is $\Theta = \{(\mu, \sigma^2) : \mu \in (-\infty, \infty), \sigma^2 \in (0, \infty)\}$;

The parameter space associated with $\mathcal{H}_0$ is $\Theta_0 = \{(\mu_0, \sigma^2) : \sigma^2 \in (0, \infty)\}$.

The likelihood function of parameter (μ, σ^2) is

$$\begin{aligned} L(\mu, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{1}{2\sigma^2}(x_i - \mu)^2\} \\ &= \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n}(x_i - \mu)^2\}. \end{aligned}$$

Check

$$\begin{aligned} \sup_{\Theta_0} L(\mu, \sigma^2) &= L(\mu_0, \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_0)^2), \\ \sup_{\Theta} L(\mu, \sigma^2) &= L(\bar{x}, \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2). \end{aligned}$$

The likelihood ratio is given by

$$\begin{aligned} \Lambda &\triangleq \frac{\sup_{\Theta_0} L(\mu, \sigma^2)}{\sup_{\Theta} L(\mu, \sigma^2)} \\ &= \frac{L(\mu_0, \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_0)^2)}{L(\bar{x}, \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2)} \\ &= \cdots\cdots \\ &= \left\{\frac{\sum_{i=1}^{n}(x_i - \mu_0)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right\}^{-n/2} \\ &= \left\{1 + \frac{n(\bar{x} - \mu_0)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right\}^{-n/2} \\ &= \left\{1 + \frac{T^2}{n-1}\right\}^{-n/2}, \end{aligned}$$

which is a decreasing function of $|T|$. Thus, a larger $|T|$ would lead us to conclude $\mathcal{H}_a$.

A $100(1-\alpha)\%$ confidence interval for μ,

$$(\bar{X} - qt(1-\alpha/2, n-1) \cdot \frac{S}{\sqrt{n}},\ \bar{X} + qt(1-\alpha/2, n-1) \cdot \frac{S}{\sqrt{n}})$$

Imagine a confidence interval

```
> x1 <- rnorm(7,3,2)
> x2 <- rnorm(3,3,2)
> z <- c(x1, x2)

> mean(z)-qt(1-0.1/2, length(z)-1)*sd(z)/sqrt(length(z))


> mean(z)+qt(1-0.1/2, length(z)-1)*sd(z)/sqrt(length(z))
```

1.4 Two samples

1.4.1 Two sample paired t-test

Assumptions:

(i) Two paired samples $(X_1, Y_1), \ldots, (X_n, Y_n)$, $EX = \mu_1,\ EY = \mu_2$;

(ii) $X_1 - Y_1, \ldots, X_n - Y_n$ are independent and normally distributed random variables with s.d. σ;

(iii) σ^2 is unknown.

Hypothesis testing problem:

$\mathcal{H}_0 :\ \mu_1 = \mu_2,$

$\mathcal{H}_a :\ \mu_1 \neq \mu_2$ (two-sided) or $\mu_1 > \mu_2$ or $\mu_1 < \mu_2$ (one-sided).

Paired Student's t-test:

Consider $d_i = X_i - Y_i,\ i = 1, \ldots, n$, and then use one sample t-test.

R codes for two sample paired t-test:

```
> t.test(x, y, alternative="two.sided", mu=0, paired=TRUE)

> t.test(x, y, alternative="greater", mu=0, paired=TRUE)

> t.test(x, y, alternative="less", mu=0, paired=TRUE)
```

1.4.2 Two samples, equal variances

Assumptions:

(i) Two independent samples $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$;

(ii) $X_1 \ldots, X_{n_1}$ is a random sample from $\mathcal{N}(\mu_1, \sigma_1^2)$;

$Y_1 \ldots, Y_{n_2}$ is a random sample from $\mathcal{N}(\mu_2, \sigma_2^2)$;

(iii) $\sigma_1^2 = \sigma_2^2$, but unknown.

Hypothesis testing problem:

$\mathcal{H}_0 : \ \mu_1 = \mu_2,$

$\mathcal{H}_a : \ \mu_1 \neq \mu_2$ (two-sided) or $\mu_1 > \mu_2$ or $\mu_1 < \mu_2$ (one-sided).

Student's t-test:

$$T \ \triangleq \ \frac{\bar{X} - \bar{Y}}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

where

$$S_p^2 \ \triangleq \ \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}, \qquad \text{(the pooled variance)}.$$

When $\mathcal{H}_0$ is true, $T \sim t_{n_1+n_2-2}$.

R codes for two sample t-test:

```
> t.test(x, y, alternative="two-sided", mu=0, var.equal=TRUE)

> t.test(x, y, alternative="greater", mu=0, var.equal=TRUE)

> t.test(x, y, alternative="less", mu=0, var.equal=TRUE)
```

For confidence intervals of $\mu_1 - \mu_2$, a pivotal quantity is

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

What is its distribution?

A two-sided $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is

$$\bar{X} - \bar{Y} \pm qt \cdot S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

(How to find `qt` ?)

1.4.3 Two samples, unequal variances

Assumptions:

(i) Two independent samples $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$;

(ii) $X_1 \ldots, X_{n_1}$ is a random sample from $\mathcal{N}(\mu_1, \sigma_1^2)$;

$Y_1 \ldots, Y_{n_2}$ is a random sample from $\mathcal{N}(\mu_2, \sigma_2^2)$;

(iii) $\sigma_1^2 \neq \sigma_2^2$, both unknown.

Hypothesis testing problem: the so-called Behrens-Fisher problem

$\mathcal{H}_0 : \ \mu_1 = \mu_2,$

$\mathcal{H}_a : \ \mu_1 \neq \mu_2$ (two-sided) or $\mu_1 > \mu_2$ or $\mu_1 < \mu_2$ (one-sided).

Welch approximate t-test:

$$T \ \triangleq \ \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}.$$

When $\mathcal{H}_0$ is true, T has a good approximate $t \sim t_\nu$, where

$$\nu \ \triangleq \ \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\left(\frac{S_1^2}{n_1}\right)^2/(n_1 - 1) + \left(\frac{S_2^2}{n_2}\right)^2/(n_2 - 1)}, \qquad \text{(an approximate number of degrees of freedom)}.$$

R codes:

```
> t.test(x, y, var.equal=FALSE, conf.level=1-alpha)
```

```
> t.test(x, y, var.equal=FALSE, conf.level=1-alpha)
```

Intuitive idea

The null hypothesis $\mathcal{H}_0 : \mu_1 = \mu_2$ says that two populations have the same mean. If this is the case, two sample means $\bar{x}$ and $\bar{y}$ have to be close enough, because `the sample is drawn from the population`. Equivalently, $\bar{x} - \bar{y}$ or T has to be close to 0 enough. The question is then: how close is "close enough "?

Consider the variance of $\bar{X} - \bar{Y}$.

$$\begin{aligned} Var(\bar{X} - \bar{Y}) &= Var(\bar{X}) + Var(\bar{Y}) \quad \text{(why?)} \\ &= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}. \end{aligned}$$

Now the question is that neither σ_1^2 or σ_2^2 is known. By the `plug-in principle`, we estimate $Var(\bar{X} - \bar{Y})$ by $\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}$, and form the Welch approximate t-test. The reason why it is an *approximate t-test* is because the exact distribution of T is unknown, and when $\mathcal{H}_0$ is true, it has a good approximation t_ν.

1.4.4 Equality of variances

Assumptions:

(i) Two independent samples $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$;

(ii) $X_1 \ldots, X_{n_1}$ is a random sample from $\mathcal{N}(\mu_1, \sigma_1^2)$;

$Y_1 \ldots, Y_{n_2}$ is a random sample from $\mathcal{N}(\mu_2, \sigma_2^2)$;

(iii) Both μ_1 and μ_2 are unknown.

Hypothesis testing problem:

$$\mathcal{H}_0 : \sigma_1^2 = \sigma_2^2 \qquad \longleftrightarrow \qquad \mathcal{H}_a : \sigma_1^2 \neq \sigma_2^2 \qquad \text{(two-sided)}$$

F test :

$$F \triangleq \frac{S_1^2}{S_2^2}$$

When $\mathcal{H}_0$ is true, $F \sim F_{n_1-1, n_2-1}$.

Reject $\mathcal{H}_0$ if the observed F value is too much greater than 1 or too much less than 1.

Derive F test via the likelihood-ratio-test method

The whole parameter space in this case is $\Theta = \{(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) : \mu_1, \mu_2 \in (-\infty, \infty), \sigma_1^2, \sigma_2^2 \in (0, \infty)\}$;

The parameter space associated with $\mathcal{H}_0$ is $\Theta_0 = \{(\mu_1, \mu_2, \sigma^2, \sigma^2) : \mu_1, \mu_2 \in (-\infty, \infty), \sigma^2 \in (0, \infty)\}$.

The likelihood function of parameter $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$ is

$$\begin{aligned} L(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) &= \prod_{i=1}^{n_1} \frac{1}{\sqrt{2\pi}\sigma_1} \exp\{-\frac{1}{2\sigma_1^2}(x_i - \mu_1)^2\} \cdot \prod_{i=1}^{n_2} \frac{1}{\sqrt{2\pi}\sigma_2} \exp\{-\frac{1}{2\sigma_2^2}(y_i - \mu_2)^2\} \\ &= \frac{1}{(\sqrt{2\pi}\sigma_1)^{n_1}(\sqrt{2\pi}\sigma_2)^{n_2}} \exp\{-\frac{1}{2\sigma_1^2}\sum_{i=1}^{n_1}(x_i - \mu_1)^2 - \frac{1}{2\sigma_2^2}\sum_{i=1}^{n_2}(y_i - \mu_2)^2\}. \end{aligned}$$

Check

$$\begin{aligned} \sup_{\Theta_0} L(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) &= L(\bar{x},\ \bar{y},\ \frac{n_1 + n_2 - 2}{n_1 + n_2} S_p^2,\ \frac{n_1 + n_2 - 2}{n_1 + n_2} S_p^2), \\ \sup_{\Theta} L(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) &= L(\bar{x},\ \bar{y},\ \frac{n_1 - 1}{n_1} S_1^2,\ \frac{n_2 - 1}{n_2} S_2^2), \end{aligned}$$

where S_p^2 is the pooled variance.

The likelihood ratio is given by

$$\begin{aligned} \Lambda &\triangleq \frac{\sup_{\Theta_0} L(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)}{\sup_{\Theta} L(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)} \\ &= \cdots\cdots \\ &= constant \cdot \left\{1 - \frac{n_2 - 1}{(n_1 - 1)F + (n_2 - 1)}\right\}^{\frac{n_1}{2}} \cdot \left\{\frac{1}{(n_1 - 1)F + (n_2 - 1)}\right\}^{\frac{n_2}{2}} \end{aligned}$$

which is a function of F.

A two-step procedure

When testing for equality of the means in two normal samples, there is a two-step procedure.

Step 1. A preliminary F-test for testing equality of the variances.

Step 2. Perform the Student's t test if the F test does not reject and equality of the variances is acceptable. Otherwise, use the Welch approximate t-test.

Example Generate 20 normal random numbers with mean=7 and sd=2, and another 30 normal random numbers with mean=5 and sd=3. Test

$$\mathcal{H}_0 : \mu_1 = \mu_2 \qquad \longleftrightarrow \qquad \mathcal{H}_a : \mu_1 \neq \mu_2 \qquad \text{(two-sided)}$$

```
> x <- rnorm(20, 7, 2)
> y <- rnorm(30, 5, 3)

                              # Step 1
> F <- var(x)/var(y)
  [1] 0.3933189               # observed F value
> pf(F, 20, 30)
  [1] 0.01642092              # p-value
                              # What are you going to do next?

                              # Step 2
> t.test(x, y,  var.equal=FALSE)


    Welch Two Sample t-test

    data:  x and y
    t = 3.4243, df = 47.877, p-value = 0.001273
    alternative hypothesis: true difference in means is not equal to 0
    95 percent confidence interval:
    0.8303074 3.1924898
```